Why Large Language Models Hallucinate?

discussion by Julio Bonis Sanz and Stuart Winter-Tear , 5/11/2025

The right term for "LLM hallucinations" is not "hallucination" but "confabulation".  
  
Hallucination (in neuroscientific terms) implies a perception (interpreted by the brain as coming from sensory organ) being created by the brain itself, not the sensory organ and thus not backed by real physical phenomena. Ex. you "hear" a voice that doesn’t exist. you "see" objects that do not exist. It is what occurs to your brain when you consume "hallucinogens" or you suffer from certain conditions (as schizophrenia or alcoholic deprivation or severe hyponatremia).  
  
Confabulation, however, is when the human brain have memory gaps (because real memories have been lost or never existed at all) and during reconstruction brain fills the gap with "fake memories". It is what happens to someone with Alzheimer when you ask them what have they eaten one hour ago and they don’t remember but just "invent" or what happens when you are asked in a trial is the killer was wearing a red or blue hat and your brain creates a memory of the assessing wearing a hat even if he was not wearing one.  
  
Once you use the exact term, so you stop saying "LLM hallucination" and start saying "LLM confabulation" it is clear that confabulations are precisely what you will expect from a system trained by masking parts of a sequence and asking the neural network to "confabulate" it.  
  
Because that is the origin of Transformer architecture and the task it was invented for: "Confabulate" the most probably translation of a sentence in English to a sentence in German after being asked to confabulate masked words in paired English-German sentences translated by humans.  
  
So transformers (LLMs) are models TRAINED TO CONFABULATE in a way that is as close as possible to what you would found in internet (training) and as close as possible as to what a low-paid human would consider a good answer (RLHF). They are confabulation machines by design!

A screenshot of a computer

Description automatically generated

# References

[1] [Do Language Models Hallucinate Because We Punish Resistance? Izzat Imran, May 10, 2025, Linkedin](https://www.linkedin.com/pulse/do-language-models-hallucinate-because-we-punish-resistance-imran-jcekc/)